High-Voltage Power Delivery Through Charge Recycling

Saravanan Rajapandian, *Member, IEEE*, Kenneth L. Shepard, *Senior Member, IEEE*, Peter Hazucha, *Member, IEEE*, and Tanay Karnik, *Senior Member, IEEE*

***Abstract*—In this paper, we describe a technique for delivering power to a digital integrated circuit at high voltages, reducing current demands and easing requirements on power-ground network impedances. The design approach consists of stacking CMOS logic domains to operate from a voltage supply that is a multiple of the nominal supply voltage. DC–DC downconversion is performed using charge recycling without the need for explicit downconverters. Experimental results are presented for the prototype system in a 0.18-μm CMOS technology operating at both 3.6 V and 5.4 V. Peak energy efficiencies as high as 93% are demonstrated at 3.6 V.**

***Index Terms*—DC–DC conversion, power delivery, power management.**

I. Introduction

OPERATING at supply voltages below 1 V, 90-nm (and below) technologies still demand in excess of 100 W of power in the largest chips, such as high-performance microprocessors. With technology scaling and increasing performance requirements, the power levels to the chip are increasing while the supply voltages are decreasing. This leads to a rapid increase in the supply current requirements. With increasing current transients and average current levels, more on-chip decoupling capacitance is required while the resistance and inductance of the power distribution network (including on-chip wiring, pins, sockets and connectors) must be kept stringently low for supply integrity. High current requirements also exacerbate on-chip electromigration concerns [1]. In this paper, we describe an energy-efficient technique for delivering power at a supply that is an integer multiple (n) of the nominal supply voltage $(V_{\rm DD})$ and which is then implicitly downcoverted to the required supply, reducing the off-chip and on-chip current requirements by a factor of n.

A high-voltage power delivery approach would traditionally require an (ideally on-chip) explicit DC–DC converter as shown in Fig. 1(a) for the case of a $2V_{\rm DD}$ external supply [2]. The most efficient DC–DC converters are buck-type regulators which generate a reduced DC level by filtering a pulse-width modulated (PWM) signal through a simple LC filter [3]. By varying the frequency or duty-cycle of the PWM signal, different DC levels can be generated. While buck converters can operate at very high efficiencies (>80%), they require off-chip inductors (to achieve required inductor sizes and Q) and occupy large area on-chip due to power transistor and filter capacitors which limits their practicality. Recent work has considered the prospect of integrating inductors on chip that include magnetic materials [4], but this clearly adds considerable cost and complexity to fabrication.

Linear regulators and switched capacitor power supplies can also be used for this downconversion but offer comparatively poor efficiencies for $nV_{\rm DD}$ to $V_{\rm DD}$ downconversion. A linear regulator consists of a power transistor whose gate voltage is controlled by a feedback error amplifier which keeps the output voltage a constant by adjusting the current through the power transistor to meet changing load requirements [5]. The energy efficiency of such a linear regulator is limited to the ratio of the regulated voltage to the input supply voltage, which is given by $1/n$ in this application. In practice, the efficiencies are further degraded by the quiescent currents necessary for regulation. In contrast, as a type of switching regulator, switched-capacitor (SC) supplies allow one to produce lower voltages at higher efficiencies than linear regulators. SC supplies are capacitive dividers, in which the capacitors are periodically “exchanged” as they are discharged by the load current. The ideal efficiency of a SC power supply is limited by the amount of “ripple” produced at the output, which can be controlled by the frequency at which the switched-capacitor supply must run. A real switched-capacitor supply suffers additional efficiency degradation due to losses in the switches and overhead associated with generating the clocks [6]. Furthermore, both linear regulators and switched-capacitor supplies consume large on-chip areas because of the power transistors (of the linear regulator) and the capacitors (of the switched-capacitor supply).

In this paper, we achieve implicit on-chip DC–DC conversion by “stacking” logic and recycling charge from one domain to another. We previously applied this technique to dynamic voltage scaling [7]. In this paper, we apply the technique for high-voltage power delivery [8], addressing new challenges in designing circuits from devices with $V_{\rm DD}$-determined bias limits operating externally from high voltages. Logic is stacked n-high and operates at an $nV_{\rm DD}$ supply. By stacking the logic domains n high, the on-chip current demands on the power and ground networks are also reduced by a factor of n over the case

Manuscript received June 28, 2005; revised February 5, 2006. This work was supported by the DARPA/MARCO C2S2 Center, by the Semiconductor Research Corporation, by the National Science Foundation under contract CCR-0086007, and by a gift from Intel.

S. Rajapandian was with the Columbia Integrated Systems Laboratory, Department of Electrical Engineering, Columbia University, New York, NY 10027 USA. He is now with Silicon Laboratories, Austin, TX 78735 USA.

K. L. Shepard is with the Columbia Integrated Systems Laboratory, Department of Electrical Engineering, Columbia University, New York, NY 10027 USA (e-mail: shepard@cisl.columbia.edu).

P. Hazucha and T. Karnik are with Intel Corporation, Hillsboro, OR 95616 USA (e-mail: peter.hazucha@intel.com; tanay.karnik@intel.com).

Digital Object Identifier 10.1109/JSSC.2006.874314

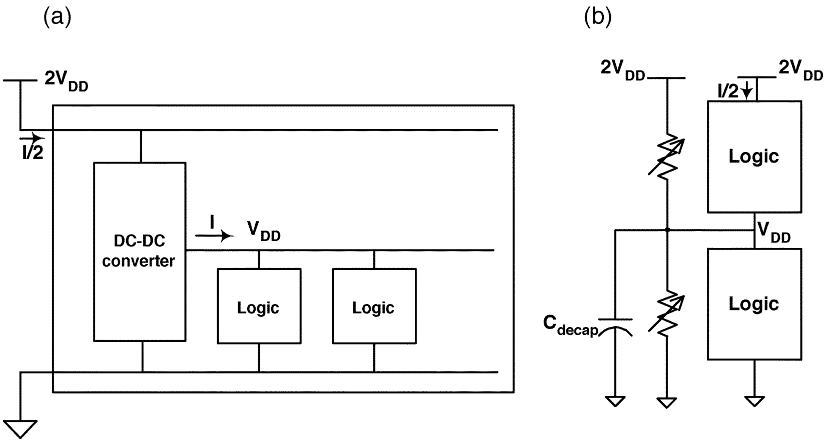


Fig. 1. (a) Explicit DC–DC conversion from $2V_{\rm DD}$ to $V_{\rm DD}$. (b) Implicit DC–DC conversion through charge recycling from $2V_{\rm DD}$ to $V_{\rm DD}$.

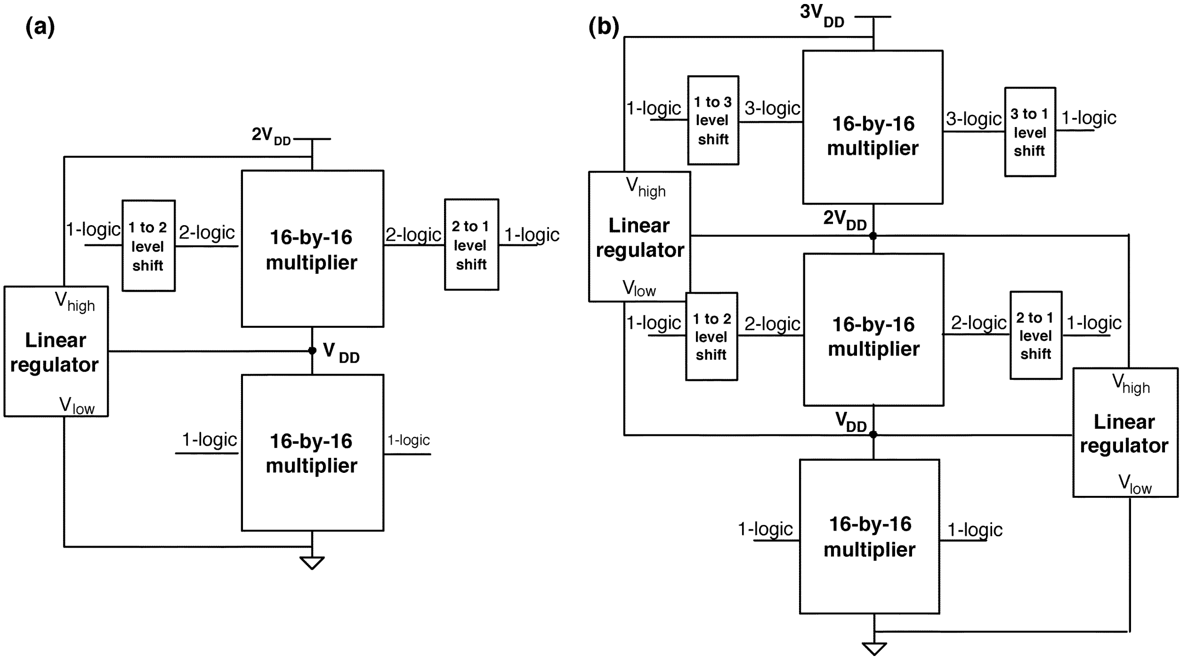


Fig. 2. (a) $2V_{\rm DD}$ system with two stacked multipliers and level shifters. (b) $3V_{\rm DD}$ system with three stacked multipliers and level shifters.

[as shown in Fig. 1(a)] of all of the domains running in parallel at $V_{\rm DD}$, which would be required with an explicit DC–DC converter. A simple two-high stack system is shown in Fig. 1(b). As the electrons drop in potential from $2V_{\rm DD}$ to $V_{\rm DD}$, the energy is consumed to perform logic in the top domain. The electrons at $V_{\rm DD}$ are recycled to perform additional logic in the bottom domain. Due to inevitable charge mismatches between the two logic domains, the internal node $V_{\rm DD}$ requires regulation, which can be achieved by the addition of a push-pull linear regulator [shown as variable resistors in Fig. 1(b)] and decoupling capacitance. This linear regulator must only source or sink the charge mismatch and, therefore, can be relatively small (and operate at lower quiescent currents) compared to linear regulators designed to accommodate the entire current of the load. High energy efficiency of the system requires careful balancing of the charge utilization of the stacked domains. Body effect issues can be avoided by using triple-well or SOI technology.

In this paper, we demonstrate the use of charge recycling DC–DC conversion to supply power to a prototype chip at both $2V_{\rm DD}$ and $3V_{\rm DD}$. In Section II, we describe the important circuit features of this prototype system. Measurement results of the testchip is discussed in detail in the Section III and Section IV concludes.

II. System Design

Our prototype system demonstrates digital logic operation at both $2V_{\rm DD}$ and $3V_{\rm DD}$ by stacking logic two- and three-high, respectively. The "stacked" logic blocks are 16-by-16 carry-save-array multipliers. In order for the stacked multipliers to communicate, we introduce level shifters as shown in Fig. 2. For example, in the $2V_{\rm DD}$ system, the multiplier in the bottom domain has logic levels between $V_{\rm DD}$ and ground and the multiplier in the top domain has logic levels between $2V_{\rm DD}$ and $V_{\rm DD}$. A linear regulator is added to the system to regulate the

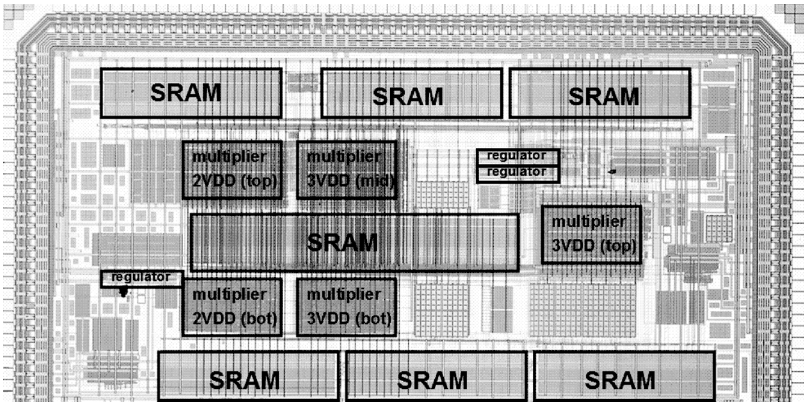


Fig. 3. Die photo of the prototype system.

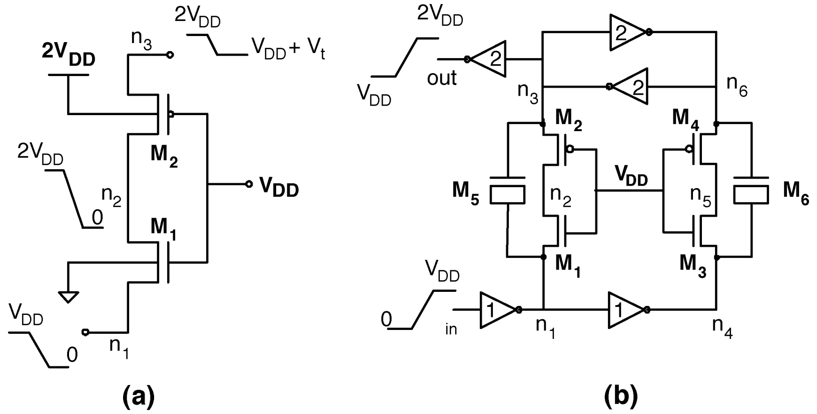


Fig. 4. (a) NMOS-PMOS stacked transistors to enable level shifting without overstressing thin oxide devices. (b) 1-to-2 shifter to convert logic levels between $V_{\rm DD}$ and 0 to logic levels between $2V_{\rm DD}$ and $V_{\rm DD}$.

internal supply nodes $V_{\rm DD}$ as shown in Fig. 2. In the $3V_{\rm DD}$ system, two linear regulators are stacked and powered as shown in Fig. 2(b) to regulate both the $2V_{\rm DD}$ and $V_{\rm DD}$ internal nodes. Both the level-shifting circuits and regulators operate over a greater-than-$V_{\rm DD}$ supply range. Despite this, thin-oxide devices are used throughout, with startup circuits to ensure that these devices are not overstressed during power-on.

The prototype system is designed and fabricated in a TSMC 0.18-μm triple-well process as shown in the die photo of Fig. 3. The system consists of stacked logic domains with level shifters, linear regulators, six-bit flash A/D converters to provide real-time monitoring of internal regulated supply voltages, and SRAMs to provide data to stacked logic domains and to store results. We now consider important circuit features of the level shifters, linear regulators, and regulated logic blocks. Startup circuitry and procedures are also explained in detail.

A. Level Shifter

Level-shifting circuits are used to interface logic levels between stacked domains. A n-to-m shifter converts a logic level between $nV_{\rm DD}$ and $(n-1)V_{\rm DD}$ (denoted as n-logic in Fig. 2) to one between $mV_{\rm DD}$ and $(m-1)V_{\rm DD}$. For example, the 1-to-2 shifter converts logic levels between $V_{\rm DD}$ and ground to logic levels between $2V_{\rm DD}$ and $V_{\rm DD}$ and is considered in Fig. 4. The NMOS-PMOS stack structure shown in Fig. 4(a) is the heart of the 1-to-2 shifter. The transistor gates are connected to $V_{\rm DD}$. When the node n_1 is at $V_{\rm DD}$, both nodes n_2 and n_3 are at $2V_{\rm DD}$. In switching the node n_1 from $V_{\rm DD}$ to ground, the nMOS transistor turns on when n_1 reaches $V_{\rm DD} - V_{\rm T}$. $M1$ pulls the node n_2 from $2V_{\rm DD}$ to ground and the node n_3 follows the node n_2 until it reaches $V_{\rm DD}+V_{\rm T}$. During this transition, the gate-source and gate-drain voltages of both transistors remain less than $V_{\rm DD}$, avoiding overstress. When the node n_3 reaches $V_{\rm DD} + V_{\rm T}$, $M2$ is off and the node n_3 is in a high-impedance state.

To prevent this high-impedance state, we use the stack structure differentially with a cross-coupled inverter pair at the output as shown in Fig. 4(b). The inverters are denoted with the number n to indicate that they are operating with n-logic signals. The inverters in the cross-coupled pair are skewed to favor the pull-down to improve switching performance. Further performance enhancement comes with the addition of MOS capacitors $M5$ and $M6$, which AC-couple nodes n_3 and n_6 to respond to switching on nodes n_1 and n_4, respectively. These capacitors, if made sufficiently large for reasonably fast slew times, also help to prevent the drain-to-source potentials of transistors $M1$ and $M3$ from exceeding $V_{\rm DD}$, reducing hot-carrier degradation issues. $M5$ and $M6$ are sized to have gate capacitances ($\cong$ 400 fF) 40 times larger than the gate capacitances of $M1$, $M2$, $M3$ or $M4$ ($\cong$ 10 fF) for typical ($\cong$ 100 ps) slew times. The simulated rise and fall delays of

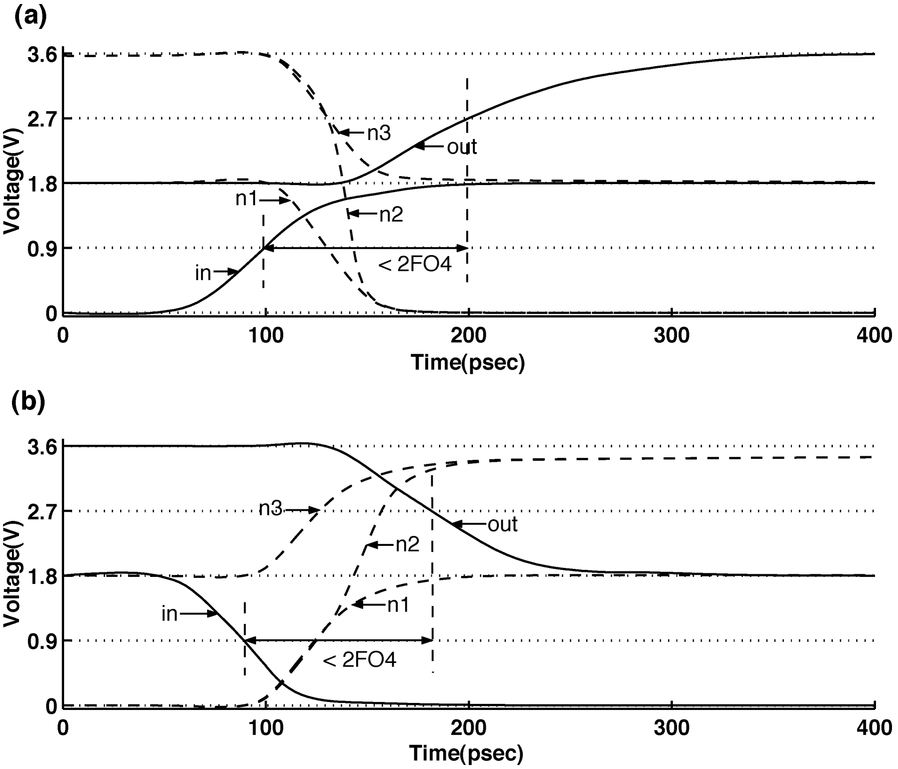


Fig. 5. (a) Simulated output-rising delay of 1-to-2 shifter. (b) Simulated output-falling delay of 1-to-2 shifter.

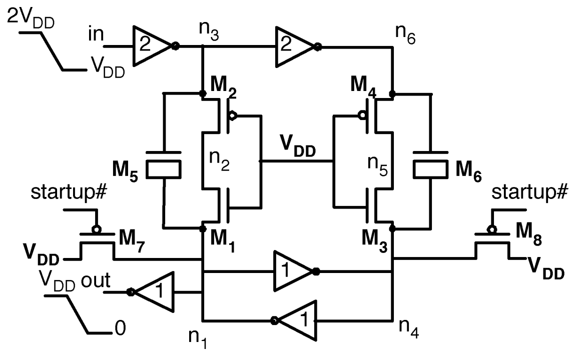


Fig. 6. 2-to-1 level shifter with startup transistors.

the 1-to-2 shifter are shown in Fig. 5. Both are less than two fanout-of-four (FO4) delays. Without the addition of transistors $M5$ and $M6$, the delays would be more than 5 FO4 delays.

The topology for the 2-to-1 shifter, shown in Fig. 6, is similar to the 1-to-2 shifter. When the input switches from $2V_{\rm DD}$ to $V_{\rm DD}$, the source of $M2$ is pulled to $2V_{\rm DD}$. The drain voltage of $M2$ follows and switches from 0 to $2V_{\rm DD}$. Transistor $M1$ pulls node n_1 toward $V_{\rm DD} - V_{\rm T}$, augmented by the action of capacitor $M5$. Once enough of a differential is established across the cross-coupled inverter pair, the output switches from $V_{\rm DD}$ to 0. In this shifter, overstress of the $M5$ and $M6$ MOS capacitors is possible during start-up. How this is avoided with a proper startup sequence and the action of devices $M7$ and $M8$ is considered below.

The 1-to-3 shifter is implemented by cascading a 1-to-2 and a 2-to-3 level shifter as shown in Fig. 7. Similarly, a 3-to-1 shifter is implemented by cascading a 3-to-2 and 2-to-1 shifter. Single-stage topologies for these shifter are also possible, as shown for the case of the 1-to-3 shifter in Fig. 7(b). The high stacks in these structure challenge switching performance and make introducing capacitors and avoiding device overstress more challenging.

B. High Speed Linear Regulator

We first consider the overall DC and transient response of a simple common-source linear regulator as shown in Fig. 8. The (closed-loop) output impedance of the linear regulator is given by

$$Z_{\rm out} = \frac{r_{o1}}{1 + A(s) g_{m1} r_{o1}} \tag{1}$$

where r_{o1} is the output resistance of saturated power transistor $M1$ (which constitutes the output impedance of the regulator in open loop) and $A(s) = A_o/(1 + s/\omega_o)$ is the gain of the error amplifier. ω_o is the pole frequency determined by the gate capacitance of the power transistor and the output impedance of the error amplifier. $A(s) g_{m1} r_{o1}$ constitutes the open-loop gain of the cascaded error amplifier and power transistor. Ignoring C_L, the output impedance of the regulator is given by

$$Z_{\rm out} = \frac{r_{o1}\left(1 + \frac{s}{\omega_o}\right)}{(1 + A_o g_{m1} r_{o1})\left(1 + \frac{s}{\omega_o(1 + A_o g_{m1} r_{o1})}\right)}. \tag{2}$$

The magnitude of this output impedance is plotted as a function of frequency in the solid curve of Fig. 9. At low frequencies, the output impedance is given by $r_{o1}/(1 + A_o g_{m1} r_{o1})$, increasing

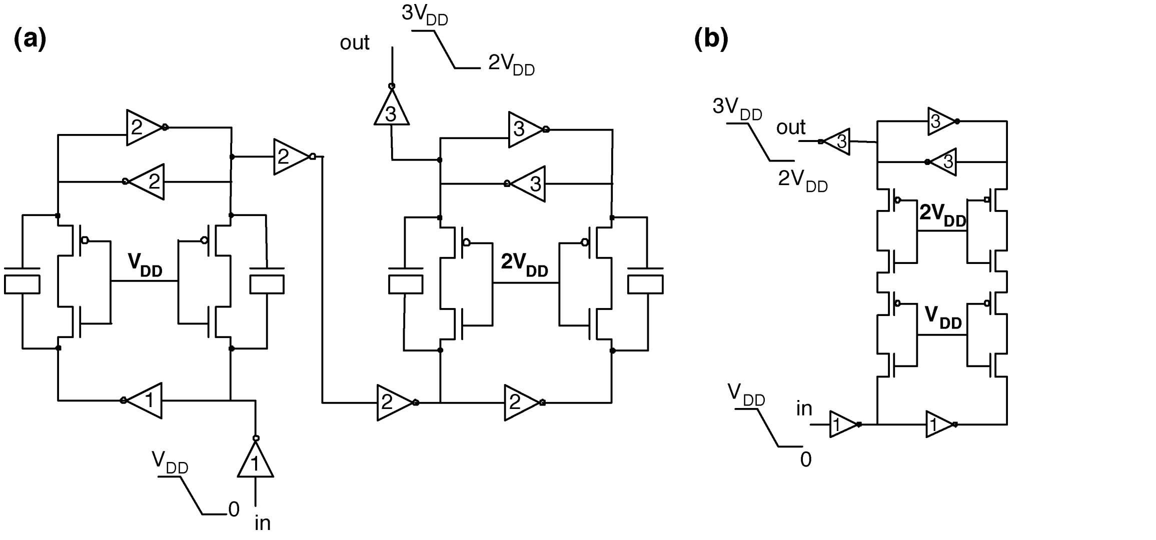


Fig. 7. (a) 1-to-3 level shifter constructed by cascading two 1-to-2 shifters. (b) 1-to-3 level shifter constructed with two NMOS-PMOS transistor stacks.

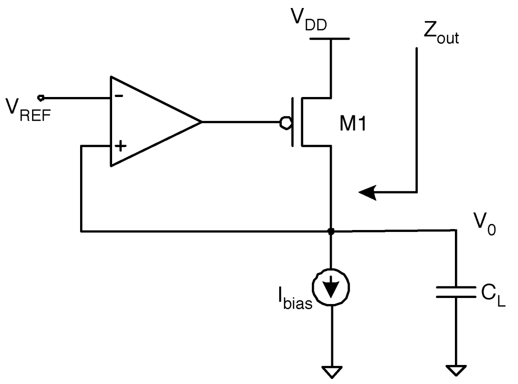


Fig. 8. Conventional linear regulator.

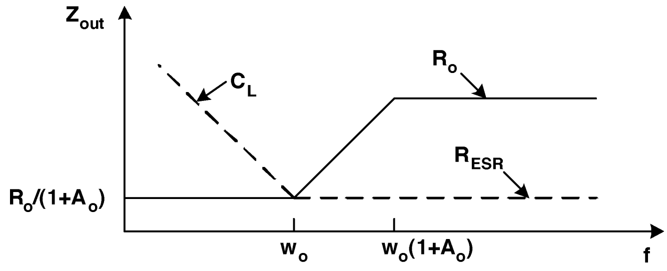


Fig. 9. Graphical illustration of the output impedance $Z_{\rm out}$ of a linear regulator, independently considering the effect of the regulation loop and the decoupling capacitor as a function of frequency.

to r_{o1} between ω_o and $\omega_o(1+A_o g_{m1} r_{o1})$ (inductive response). For frequencies below $\omega_o(1+A_o g_{m1} r_{o1})$, the linear regulator can be modeled as a constant voltage source with series resistor (R_s) and inductor (L_s) given by

$$R_s = \frac{r_{o1}}{1+A_o g_{m1} r_{o1}} \tag{3}$$

$$L_s = \frac{r_{o1}}{\omega_o(1+A_o g_{m1} r_{o1})}. \tag{4}$$

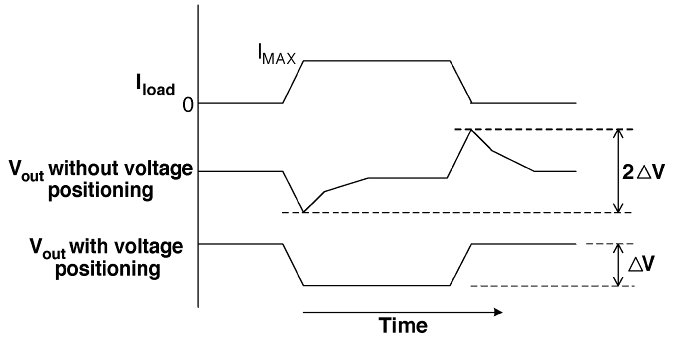


Fig. 10. Optimum transient response with voltage positioning.

This simple model ignores other high-frequency poles and zeros associated with the parasitics of the power transistor.

To reduce the output impedance at high-frequencies, a decoupling capacitance C_L with effective series resistance $R_{\rm ESR}$ is added in parallel with the regulator as shown in Fig. 9. The magnitude of impedance of the series resistor-capacitor network is shown as the dotted line in Fig. 9, becoming $R_{\rm ESR}$ at high frequencies. When a worst-case current step of magnitude $I_{\rm MAX}$ is applied at the output, the output drops by ΔV over a time interval T_R $(= C_L \Delta V / I_{\rm MAX})$ for a decoupling capacitance C_L. T_R, consequently, determines the required response time of the regulator to achieve this ΔV target for the given amount of decoupling capacitance. A high open loop bandwidth (ω_o) for the regulator reduces the decoupling capacitor required to maintain a given output impedance, improving this response time.

Optimal droop response is achieved when the output impedance of the regulator is resistive over a wide frequency range of the load current spectrum, a response known as "voltage positioning" [9], [10]. As shown in Fig. 10, a regulator without voltage positioning gives an output droop of ΔV when the load current changes from 0 to $I_{\rm MAX}$ and an overshoot of ΔV when load current goes back to 0. The total output transient is 2ΔV. In contrast, a regulator with voltage positioning has output transient of only ΔV for the same load current transient because the DC and AC impedances are comparable. The

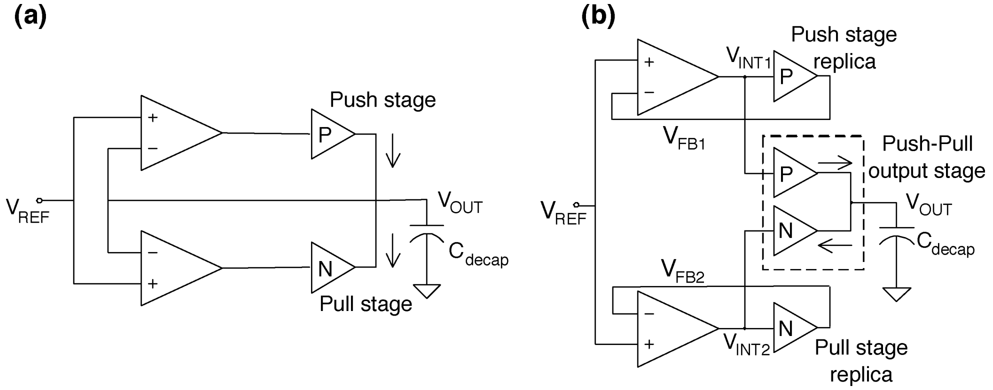


Fig. 11. (a) Single-loop push-pull linear regulator and (b) dual-loop push-pull linear regulator.

condition to establish this nearly constant impedance for the regulator of Fig. 8 is given by

$$R_{\text{ESR}} = \frac{R_o}{1 + A_o} \tag{5}$$

$$\frac{L}{R_o/(1 + A_o)} = C_L R_{\text{ESR}}. \tag{6}$$

The linear regulators as employed in this charge-recycling implicit voltage downconversion approach are designed to compensate for charge *mismatches* between stacked domains, providing load regulation of the internal supply nodes in the presence of what may be rapid changes in net current flow into these supply nodes. Fig. 11(a) shows a conventional push-pull linear regulator with single loop controlling the push (P) and pull (N) stages. Fig. 11(b) shows the dual-loop push-pull voltage regulator used in this design, in which push and pull stage replicas are used in one set of two feedback loops to force the output voltage to track V_{ref}. In the case of large open-loop gains for all the feedback loops, DC values of V_{OUT}, V_{FB1}, and V_{FB2} are all set by V_{ref}. A second set of two negative feedback loops within the push-pull output stages respond to load current transients without requisite changes in V_{INT1} and V_{INT2} and provide a low output impedance to high bandwidths. The push-pull stages (and their replicas) are noninverting.

Fig. 12 shows the detailed implementation of the push-pull linear regulators used in this design and how they are stacked to regulate both $2V_{\text{DD}}$ and V_{DD} internal supplies for a $3V_{\text{DD}}$ external reference. Feedback loops 1 and 2 set V_{out} according to V_{ref}. Feedback loops 3 and 4 respond to load current transients. The error amplifiers are single-stage cascoded differential pairs. The replica stages from V_{INT1} to V_{FB1} and from V_{INT2} to V_{FB2} have approximately unity gain. The overall loop gain of feedback loops 1 and 2 is 80 dB with a unity-gain bandwidth of only 50 MHz. These gain-bandwidth targets allow V_{FB1} and V_{FB2} to accurately track V_{ref} but at low enough current levels so as not to significantly contribute to quiescent current. The error amplifiers are biased with 100 μA each, as are the replica push and pull stages. The output push and pull stages are 30 times the size of their replicas and self-biased with a quiescent current of 3 mA to provide faster response time in these loops. This leads to a matched voltage biasing of the devices in the output stage and replica.

In operation, if the load sinks current the pull (N) stage is nearly off and the common-gate feedback loop 3 acts to source the required load current. Similarly, if the load sources current, the push (P) stage is nearly off and the common-gate feedback loop 4 acts to sink the load current. Voltage changes at the output produced by load current transients are amplified by transistors $M1$ and $M4$ acting as common gate amplifiers with open-loop gains of approximately 15. Ten-percent output voltage droops translate into large power transistor overdrives. Transistor $M3$ has a relatively large V_{gs} swing given by $V_{gs3} \cong V_{\text{DD}} - (V_{gs1} - V_{T1})$. A similarly large swing on V_{gs6} allows both power transistors ($M3$ and $M6$) to be modestly sized and still source or sink significant load current. The high voltage headroom in these regulators allows these devices to operate with high overdrive while still remaining saturated. Both the push and pull stage are sized to source or sink 40 mA (I_{MAX}) of current with a five-percent (90 mV) droop requirement, leading to an output impedance requirement of 2.25 Ω. The regulator consumes a quiescent current of 4 mA, leading to a 90% current efficiency at I_{MAX}. Any noise in the output voltage V_{OUT} can couple through the gate-to-source capacitance of $M1$ and $M4$ and affect the stability of the fairly low bandwidth loops 1 and 2. To prevent this, capacitors $C1$ and $C2$, each 4 pF, are added to stabilize the loops and to provide constant gate bias to transistors $M1$ and $M4$.

Calculating the output impedance of the push stage follows the analysis of the generalized regulator of Fig. 8. The common gate amplifier $M1$ acts as the error amplifier with a gain of $g_{m1}r_{o2}$, where g_{m1} is the small signal transconductance of $M1$ and r_{o2} is the output impedance of transistor $M2$. The open-loop gain can be calculated by breaking the loop at the source of transistor $M1$ and by adding the impedance seen by transistor $M3$ at its drain. $M1$ and $M2$ are sized such that $g_{m1}R_L \gg 1$ and $r_{o1} \gg r_{o2}$, where R_L is the linearized resistance of the load. In this case, the loop gain of the output stage is given by $g_{m3}r_{o2}$, where g_{m3} is the small signal transconductance of $M3$. The internal pole associated with the output stage is given by

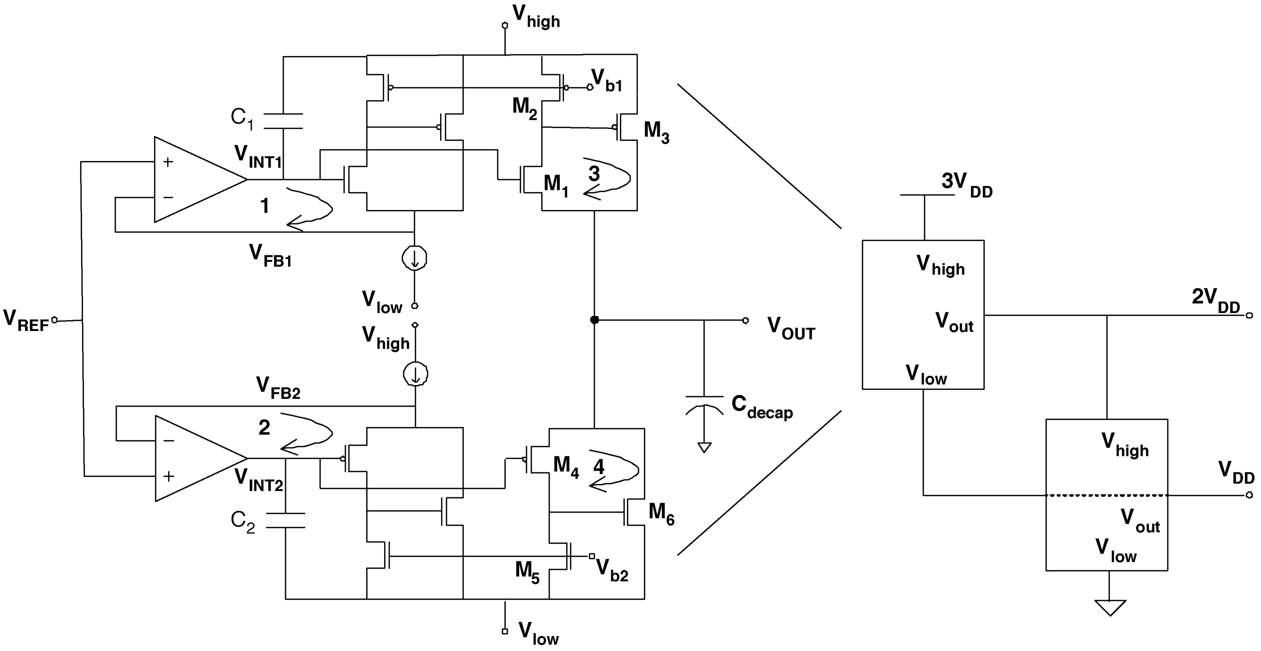


Fig. 12. Dual loop push-pull linear regulator.

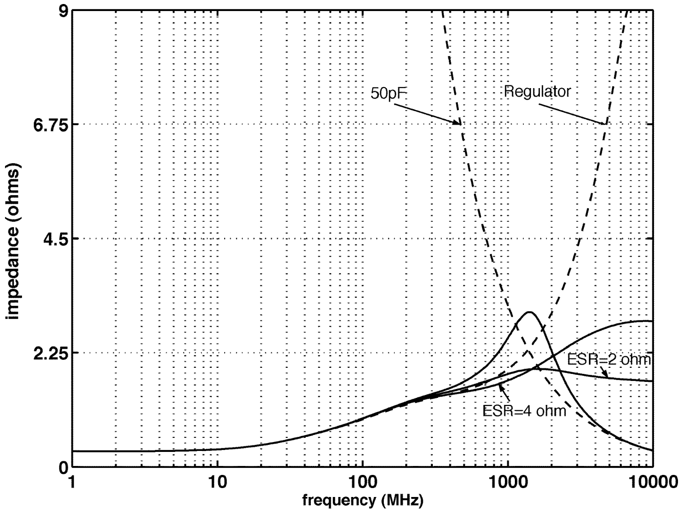


Fig. 13. Output impedance of the push-pull regulator.

$1/C_{gs3}r_{o2}$ where C_{gs3} is the gate to source capacitance of M3. The output impedance of the push stage is given by

$$Z_{\text{out}} = \frac{1/g_{m1}(1+sC_{gs3}r_{o2})}{1+g_{m3}r_{o2}}. \tag{7}$$

We note that linear regulators based on common-source output stages traditionally have a dominant pole determined by the gate capacitance of the power transistor [11], [12]. This leaves the load capacitance as a secondary pole, leading to stability challenges. The modest sizes of $M3$ and $M6$, enabled by their large V_{gs} swings and the significant headroom from regulating voltages V_{DD} from the rail, allows the time constants associated with the gate capacitances to be pushed to high frequencies, allowing the output decoupling capacitance (C_L) to act as the dominant pole. Stability can be achieved with a modest value of $C_L = 50$ pF,[1] improving regulator bandwidth and response time. For $\Delta V = 180$ mV the response time of the regulator is 225 ps.

R_{ESR} of the explicit thin-oxide capacitor is determined by its channel length. For voltage positioning response, the decoupling capacitance and effective series resistance must be chosen to extend the DC output impedance to high frequencies. Fig. 13 shows simulation results for the impedance of the push-pull linear regulator in the absence of additional decoupling capacitance, the impedance of the decoupling capacitance $(C_L = 50$ pF) alone, and the combined impedance of the linear regulator with decoupling capacitance. Three different values of R_{ESR} are considered: 0 Ω, 2 Ω, and 4 Ω. The linear regulator has a inductive response for frequencies greater than 1 GHz, reduced by C_L at high frequencies at which point it is determined by R_{ESR}. An R_{ESR} value of 2 Ω delivers the "flattest" impedance response.[2] The linear regulator is implemented using only thin-oxide devices with no overstress issues under steady-state operation. Start-up requires special consideration and is discussed below.

It can be shown the control of the quiescent current in this dual loop regulator [of Fig. 11(b)] is significantly better than this control in the single-loop push-pull topology [of Fig. 11(a)] [7]. Consider the push and pull stages individually; that is, before they are connected together to drive the common output V_{out}.

[1] In addition to this explicit thin-oxide capacitors, implicit decoupling capacitance is also present due to well capacitance and nonswitching gate capacitance of transistors in the multiplier. We estimate this to be approximately 10 pF on the $2V_{\text{DD}}$ node and 14 pF on the V_{DD} node.

[2] We attribute the frequency response evident in Fig. 13 in the 10–300 MHz range to pole-zero doublets introduced by the coupling of the output to nodes V_{INT1} and V_{INT2} through the C_{gs} of devices $M1$ and $M4$. The impedeances at nodes V_{INT1} and V_{INT2} due to the amplifiers of loops 1 and 2 are inductive around 50 MHz. This impedance is in parallel with the capacitance on nodes V_{INT1} and V_{INT2}.

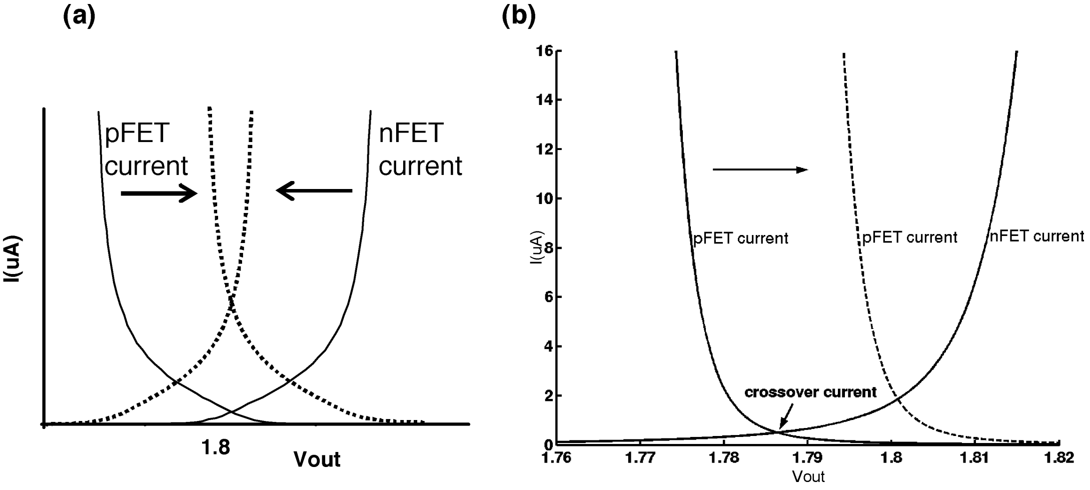


Fig. 14. (a) Typical power transistor current of a single loop push-pull linear regulator as a function of output voltage. (b) Simulated power transistor current of a dual-loop push-pull linear regulator as a function of output voltage.

We further assume that the output stages are biased to match the replicas. Feedback loops 1 and 2 enable the output voltages of the two stages to track V_{FB1} and V_{FB2}, respectively. When these two output stages are connected to drive the same output V_{out}, V_{out} settles to a voltage between V_{FB1} and V_{FB2}, the value of which depends on the relative output impedance of the push and the pull stages. When $V_{\mathrm{FB1}} = V_{\mathrm{FB2}} = V_{\mathrm{REF}}$, a nominal quiescent current flows through the output stage. Any static difference between V_{FB1} and V_{FB2}, however, results in additional quiescent current as determined by the output impedance of the push and pull stages. The nominal currents through transistors $M3$ and $M6$ are plotted as a function of the output voltage in the solid curves of Fig. 14(b). Where these curves cross determines the quiescent current. An offset voltage, for example, in the error amplifier of loop 1 results in a higher value of V_{int1}, shifting the pFET curve as shown in the dotted curve of Fig. 14(b). The result is a relatively insignificant increase in the quiescent current. This contrasts sharply with the single-loop system of Fig. 14(a). In this case, offset in any of the opamps results in shifting of both pFET curve and nFET curve. The result can be a relatively significant increase in the quiescent current [13]. Any mismatch between the push-pull stage and its replica circuit can be modelled as offset voltages at the inputs of the error amplifiers. These mismatches are minimized by placing the push-pull and replica stages in close proximity.

C. Digital Blocks

The "stacked" digital blocks chosen for our prototype (as shown in Fig. 2) are 16-by-16 pipelined in Fig. 2) are 16-by-16 pipelined fixed-point multipliers designed to operate at 650 MHz, synthesized into a digital standard cell library in a triple-well TSMC 0.18-μm process. Activity of the blocks is controlled through data dependence in which a four-bit control is used to zero out the input bits of each of the operands of the multiplier. The four bits are decoded to appropriate operand size with the remaining bits zeroed out. For example, when the control word is "1111", two 16-bit numbers are multiplied, while when the control word is "1000", the operands are reduced to nine bits. The multiplier is designed to consume a maximum power of 110 mW (60 mA of load current) at 650 MHz. I_{MAX} for the regulator is, therefore, 66% of the maximum current that can be sourced or sunk by the logic domains. All the inputs of the multiplier are sourced of "upconverting" level shifters while all the outputs feed "downconverting" level shifters. It is important to note that in this "stacking" technique only well–substrate junctions, which have breakdown voltage in this technology of more than 15 V, are biased with potentials in excess of V_{DD}.

D. System Startup

Both the level shifters and the linear regulators operate over a greater-than-V_{DD} supply range. While no devices are biased with V_{GS} or V_{DS} in excess of V_{DD} in steady-state, special attention must be paid during power-on to avoid device overstress.

For example, during power-on of the 2-to-1 level shifter, as shown in Fig. 6, that is, as the external supply voltage is increased, the one-logic inverters turn on before the two-logic inverters; the former are on when the external supply reaches V_{DD}. As a result, the stable operating point for the cross-coupled inverters is established before the input differential voltage for the level shifter is set. Nodes n_1 and n_4 can power up as ground and V_{DD}, respectively, while nodes n_3 and n_6 can power up at $2V_{\mathrm{DD}}$ and V_{DD}, respectively, leaving a $2V_{\mathrm{DD}}$ bias across capacitor $M5$. To prevent this, startup switches $M7$ and $M8$ hold nodes n_1 and n_4 at V_{DD} during power on.

Similar start-up issues are associated with the linear regulators and internal supply nodes. We consider the start-up of the $2V_{\mathrm{DD}}$ system as shown in Fig. 15, although identical considerations and techniques apply to the $3V_{\mathrm{DD}}$ system. The linear regulator is not active until the external supply (V_{supply}) is powered to $2V_{\mathrm{DD}}$. An improper power-on sequence can leave the internal node unregulated at a voltage much less than V_{DD}, resulting in overstress of devices in the linear regulator and in the top voltage domain when V_{supply} reaches $2V_{\mathrm{DD}}$. To prevent this, we first increase V_{supply} to V_{DD}. The regulator is not active to regulate the internal node V_{int} at V_{DD}. As a result, the bias V_{startup} is used to set V_{int} through forward-biased diode D_1. V_{startup} is increased from 0 to $V_{\mathrm{DD}} + V_{\mathrm{D}}$, where V_{D} is the forward-biased diode drop

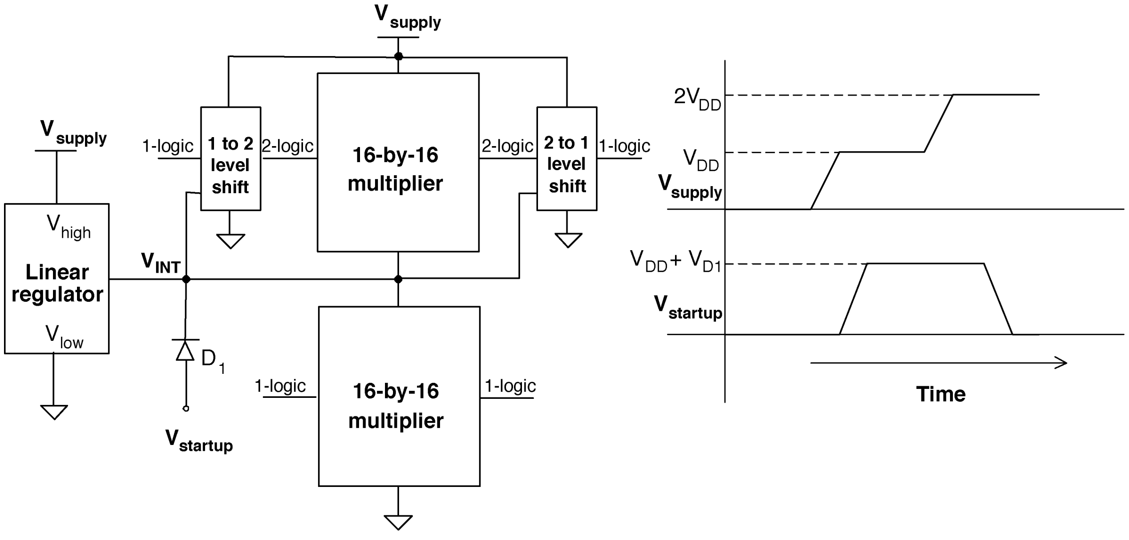


Fig. 15. Startup circuitry with timing diagram to power up the system.

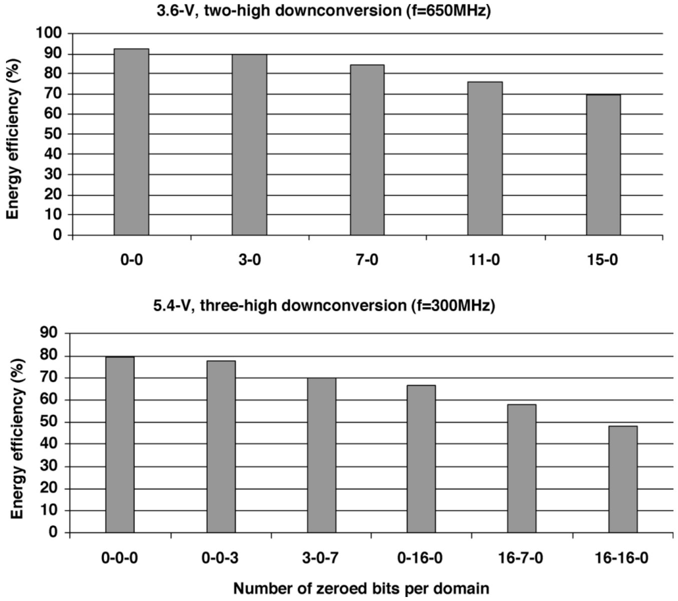


Fig. 16. Energy efficiency of the $2V_{\rm DD}$ and $3V_{\rm DD}$ system at 650 MHz and 300 MHz, respectively.

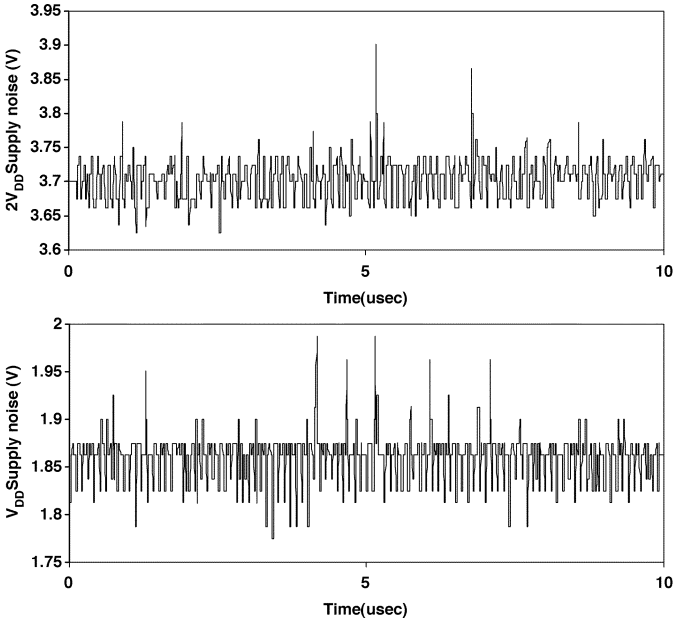


Fig. 17. 10% regulation of $2V_{\rm DD}$ and $3V_{\rm DD}$ as measured by the on-chip flash converters.

across D_1, maintaining $V_{\rm int}$ at $V_{\rm DD}$ as $V_{\rm supply}$ is increased to $2V_{\rm DD}$. Once the external supply is increased to $2V_{\rm DD}$, $V_{\rm startup}$ is brought down to ground, reverse biasing D_1.

III. Measurement Results

In Fig. 16, we show the measured energy efficiency as a function of relative activity between the domains; the numbers denote the number of *zeroed* bits in each domain (e.g., for the $3V_{\rm DD}$ system, 0-2-0 indicates that the leading two bits of the multiplier and multiplicand input to the middle domain are zeroed). Peak energy efficiency of almost 93% is achieved for the 0-0 case in the $2V_{\rm DD}$ system at 650 MHz. Results for the $3V_{\rm DD}$ system are shown at a 300-MHz operating frequency because of an unintended long path in the design, which did not allow 650-MHz operation. In this case, the current consumed by each multiplier is only 30 mA and the quiescent current in each regulator is around 4 mA resulting in low current efficiency.

The chip incorporates three on-chip 40-MHz, six-bit flash ADCs to monitor the regulation of the internal $V_{\rm DD}$ and $2V_{\rm DD}$ nodes in real-time.[3] These identical ADCs operate over two different reference voltage ranges depending on whether a $2V_{\rm DD}$ or a $V_{\rm DD}$ supply is being monitored. For example, to measure the internal node voltage $2V_{\rm DD} = 3.6$ V, the ADC operates between 4.5 V to 2.7 V. The clock to the ADC is upconverted to voltage levels between 4.5 V and 2.5 V and the digital output is level shifted from logic levels between 4.5 V to 2.7 V to logic levels between 1.8 V and ground. Fig. 17 shows the outputs

[3] For observed waveforms, aliasing will occur for frequency content in the power supply above 20 MHz. Since the purpose of these current monitors is to determine peak voltage variations, this is not a concern.

TABLE I
SUMMARY OF CHARACTERISTICS AND MEASURED PERFORMANCE

Measured peak energy efficiency	$2V_{DD}$ system (at 650 MHz)	93%
	$3V_{DD}$ system (at 300 MHz)	80%
Area	Total Area ($2V_{DD}$ system)	$0.51mm^2$
	Total Area ($3V_{DD}$ system)	$0.84mm^2$
	Single multiplier logic domain	$0.2mm^2$
	Single logic domain with level shifters	$0.27mm^2$
	Level Shifter	$0.001mm^2$
	Regulator ($2V_{DD}$-to-V_{DD})	$0.044mm^2$
	Regulator ($3V_{DD}$-to-$2V_{DD}$)	$0.053mm^2$
Pwr. Tran. Widths ($2V_{DD}$-to-V_{DD} regulator)	PMOS	$0.6mm$
	NMOS	$0.3mm$
Pwr. Tran. Widths ($3V_{DD}$-to-$2V_{DD}$ regulator)	PMOS	$1mm$
	NMOS	$0.5mm$
Regulator quiescent current	$2V_{DD}$-to-V_{DD} regulator	$4mA$
	$3V_{DD}$-to-$2V_{DD}$ regulator	$3.6mA$
Maximum load current (I_{MAX})	$2V_{DD}$-to-V_{DD} regulator	$40mA$
	$3V_{DD}$-to-$2V_{DD}$ regulator	$65mA$
Regulator response time	Output stage	225 $psec$
	Replica bias loop	20 $nsec$

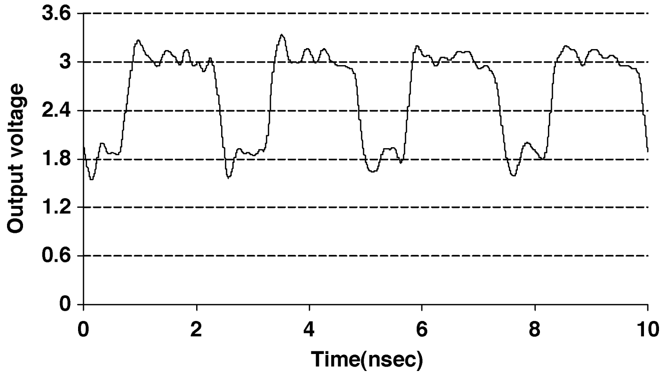


Fig. 18. Measured 1-to-2 level-shifted output of a 400-MHz ground-to-$V_{\rm DD}$ clock signal.

of these power-supply-monitoring ADCs, showing better than 10% regulation for the multipliers running with random input patterns. In Fig. 18, we show the measured transient operation of a 1-to-2 level shifter transforming the 400-MHz multipler clock input, obtained by picoprobing.[4]

Table I summarizes the important characteristics and performance of both $2V_{\rm DD}$ and $3V_{\rm DD}$ power delivery systems.

IV. CONCLUSIONS AND DISCUSSION

We have presented an approach for *stacking* logic to achieve implicit on-chip DC–DC conversion with greater than 90% energy efficiency and little area overhead. We have shown that high-voltage power delivery can be used to reduce current levels in power-ground distribution networks, easing requirements on the impedances of these networks. The high-voltage power delivery system described here uses only thin-oxide devices and includes start-up devices to avoid device overstress during power-on. High-bandwidth, high-voltage-tolerant linear regulators with bandwidths in excess of 1 GHz are designed for internal node regulation. Level shifters that translate logic levels between stacked domains are designed and implemented.

The approach clearly requires that the stacked loads have well-balanced charge utilization for high efficiency. This means that any chock gating must be balanced across the domains. One context in which this approach may be more easily applicable is in a multicore microprocessor in which each core could be designed to operated in a different stacked domain. Current utilization in each domain could be controlled with workload balancing; level-shifting voltage interfaces would only have to be present to interface between cores or with the chip pads.

ACKNOWLEDGMENT

The authors acknowledge D. Ginsberg of Columbia University (now with IBM) for helping with chip layout.

REFERENCES

[1] J. Trattles, A. O'Neill, and B. Mecrow, "Three-dimensional finite-element investigation of current crowding and peak temperatures in VLSI multilevel interconnections," *IEEE Trans. Electron Devices*, vol. 40, no. 7, pp. 1344–1347, Jul. 1993.

[2] G. Schrom *et al.*, "Feasibility of monolithic and 3D-stacked DC-DC converters for microprocessors in 90 nm technology generation," in *Proc. Int. Symp. Low Power Electronics and Design*, 2004, pp. 263–268.

[4] The ringing observed in this waveform is an artifact of the probing.

[3] G. Wei and M. Horowitz, "A fully digital, energy-efficient, adaptive power-supply regulator," *IEEE J. Solid-State Circuits*, vol. 34, pp. 1659–1671, Apr. 1999.

[4] D. Gardner, A. Crawford, and S. Wang, "High frequency (GHz) and low resistance integrated inductors using magnetic material," in *Proc. IEEE Int. Interconnect Technology Conf.*, Jun. 2001, pp. 101–103.

[5] T. Endoh, K. Sunaga, H. Sakuraba, and F. Masuoka, "An on-chip 96.5% current efficiency CMOS linear regulator using a flexible control technique of output current," *IEEE J. Solid-State Circuits*, vol. 36, no. 1, pp. 34–39, Jan. 2001.

[6] G. Patounakis, Y. W. Li, and K. L. Shepard, "A fully integrated on-chip DC-DC conversion and power management system," *IEEE J. Solid-State Circuits*, vol. 39, no. 3, pp. 443–451, Mar. 2004.

[7] S. Rajapandian, Z. Xu, and K. Shepard, "Energy-efficient low-voltage operation of digital CMOS circuits through charge-recycling," in *Int. Symp. VLSI Circuits Dig.*, Jun. 2004, pp. 330–333.

[8] S. Rajapandian, K. Shepard, P. Hazucha, and T. Karnik, "High-tension power delivery: operating 0.18 μm CMOS digital logic at 5.4 V," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig.*, Feb. 2005, pp. 298–299.

[9] A. Waizman and C. Chung, "Resonant free power network design using extended adaptive voltage positioning (EAVP) methodology," *IEEE Trans. Adv. Packag.*, vol. 24, no. 3, pp. 236–244, Aug. 2001.

[10] P. Hazucha *et al.*, "Area-efficient linear regulator with ultra-fast load regulation," *IEEE J. Solid-State Circuits*, vol. 40, no. 4, pp. 933–940, Apr. 2005.

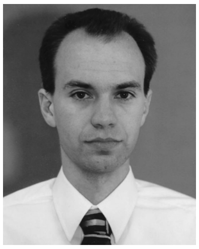
[11] L. Carley and A. Aggarwal, "A completely on-chip voltage regulation technique for low power digital circuits," in *Proc. Int. Symp. Low-Power Electronics and Design (ISLPED)*, 1999, pp. 109–111.

[12] S. Jou and T. Chen, "On-chip voltage down converter for low-power digital system," *IEEE Trans. Circuits Syst. II, Analog Digit. Signal Processing*, vol. 45, no. 5, pp. 617–625, May 1998.

[13] H. Khorramabadi, "A CMOS line driver with 80-db linearity for ISDN applications," *IEEE J. Solid-State Circuits*, vol. 27, no. 4, pp. 539–544, Apr. 1992.

Saravanan Rajapandian (S'04–M'05) received the B.Tech. degree in electrical engineering from the Indian Institute of Technology, Madras, in 2001, and the M.S. and Ph.D. degrees in electrical engineering from Columbia University, New York, NY, in 2003 and 2005, respectively.

He is currently with Silicon Laboratories, Austin, TX. His research interests include DC–DC converters, high-speed linear regulators, and low-power digital circuits.



Kenneth L. Shepard (M'92–SM'03) received the B.S.E. degree from Princeton University, Princeton, NJ, in 1987, and the M.S. and Ph.D. degrees in electrical engineering from Stanford University, Stanford, CA, in 1988 and 1992, respectively.

From 1992 to 1997, he was a Research Staff Member and Manager in the VLSI Design Department at the IBM T. J. Watson Research Center, Yorktown Heights, NY, where he was responsible for the design methodology for IBM's G4 S/390 microprocessors. Since 1997, he has been with Columbia University, where he is now Associate Professor. He also served as Chief Technology Officer of CadMOS Design Technology, San Jose, CA, until its acquisition by Cadence Design Systems in 2001. His current research interests include design tools for advanced CMOS technology, on-chip test and measurement circuitry, low-power design techniques for digital signal processing, low-power intrachip communications, and CMOS imaging applied to biological applications.

Dr. Shepard received the Fannie and John Hertz Foundation Doctoral Thesis Prize in 1992. At IBM, he received Research Division Awards in 1995 and 1997. He was also the recipient of an NSF CAREER Award in 1998 and IBM University Partnership Awards from 1998 through 2002. He was also awarded the 1999 Distinguished Faculty Teaching Award from the Columbia Engineering School Alumni Association. He has been an Associate Editor of IEEE Transactions on Very Large Scale Integration Systems and was the technical program chair and general chair for the 2002 and 2003 International Conference on Computer Design, respectively. He has served on the program committees for ICCAD, ISCAS, ISQED, GLS-VLSI, TAU, and ICCD.

Peter Hazucha (M'01) was born in Slovakia. He received the Ph.D. degree in physics from Linkoping University, Sweden, in 2000.

Since 2001, he has been with Circuit Research Laboratory at Intel. For the past five years, he has been actively working on neutron soft-error rate characterization of CMOS processes including P1262, and soft-error hardening techniques. Since 2001, he was also involved in design of high-frequency voltage regulators and DC-DC converters.

Tanay Karnik (M'88–SM'04) received the Ph.D. degree in computer engineering from the University of Illinois at Urbana-Champaign in 1995.

He is currently a Principal Engineer at Circuit Research Laboratory at Intel. From 1995 to 1999, he worked in the Strategic CAD Laboratory at Intel, working on RTL partitioning, physical design and special circuits layout. Since March 1999, he has led the power delivery, soft error rate, and optoelectronic circuits research in the Circuits Research, Intel Labs. Earlier, from 1987 to 1988, he worked on programmable logic controller design at Larsen & Toubro Ltd. in India. He spent the summer of 1994 at AT&T Bell Labs developing a timing and synthesis module for FPGAs. His research interests are in the areas of power delivery, soft errors, voltage regulator module (VRM) circuits, leakage tolerance and physical design. He has published over 30 technical papers, and has 18 issued and 47 pending patents in these areas. He has presented several invited talks and tutorials. He has also graduated 3 PhD students as an industrial advisor.

Dr. Karnik serves on the ICCAD, ISQED, DAC, and ICICDT committees. He is on the review committees of the IEEE Journal of Solid-State Circuits, IEEE Transactions on Computer-Aided Design, IEEE Transactions on Circuits and Systems, and IEEE Transactions on VLSI. He will serve as the TPC Chair for ISQED'06.